doi:10.1016/S0022-2836(02)00587-9 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 321, 7–20

JMB

Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors

Leonid A. Mirny[1]* and Mikhail S. Gelfand[2]

[1]*Harvard-MIT Division of Health Science and Technology Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 USA*

[2]*Integrated Genomics–Moscow P.O. Box 348, Moscow 117333 Russian Federation*

**Corresponding author*

Concepts of orthology and paralogy are become increasingly important as whole-genome comparison allows their identification in complete genomes. Functional specificity of proteins is assumed to be conserved among orthologs and is different among paralogs. We used this assumption to identify residues which determine specificity of protein–DNA and protein–ligand recognition. Finding such residues is crucial for understanding mechanisms of molecular recognition and for rational protein and drug design. Assuming conservation of specificity among orthologs and different specificity of paralogs, we identify residues that correlate with this grouping by specificity. The method is taking advantage of complete genomes to find multiple orthologs and paralogs. The central part of this method is a procedure to compute statistical significance of the predictions. The procedure is based on a simple statistical model of protein evolution. When applied to a large family of bacterial transcription factors, our method identified 12 residues that are presumed to determine the protein–DNA and protein–ligand recognition specificity. Structural analysis of the proteins and available experimental results strongly support our predictions. Our results suggest new experiments aimed at rational re-design of specificity in bacterial transcription factors by a minimal number of mutations.

Keywords: orthologs; paralogs; mutual information; covariation; transcription

Introduction

The concepts of orthology and paralogy were originally introduced by Walter Fitch in 1970[1,2] and recently became a subject of active discussion.[3–6] Briefly, orthologs are genes in different organisms which are direct evolutionary counterparts of each other. Orthologs were inherited through speciation, as opposed to paralogs which are genes in the same organism which evolved by gene duplication.[6,3,2] After duplication, paralogous proteins experience weaker evolutionary pressure and their specificity diverges leading to emerging of new specificities and functions. Orthologous proteins, on the contrary, are believed to be under similar regulation, have the same function and usually the same specificity in close organisms.[7–9] In other words, both paralogs and orthologs are assumed to have similar general biochemical functions, while orthologs are also believed to have the same specificity. Although the validity of these assumptions is yet to be verified experimentally, numerous case studies support such views.[6,10] Several methods have been developed to find orthologous proteins in complete genomes.[8,11] The assumption of similar regulation of orthologous proteins was productively used by several groups to identify common regulatory motifs upstream of orthologous proteins.[9,12–15] In this study we exploit another property of orthologs: similar specificity, as contrasted by different specificities of paralogs.

If the above assumption is correct, grouping by orthology becomes grouping of proteins by specificity. Here we developed a method, which uses such grouping to identify amino acid residues that determine the protein specificity. Specificity-determining residues can be very hard to find even when the structure of a protein or a complex

Abbreviations used: MSA, multiple sequence alignment; PDB, Protein Data Bank.

E-mail address of the corresponding author: leonid@mit.edu

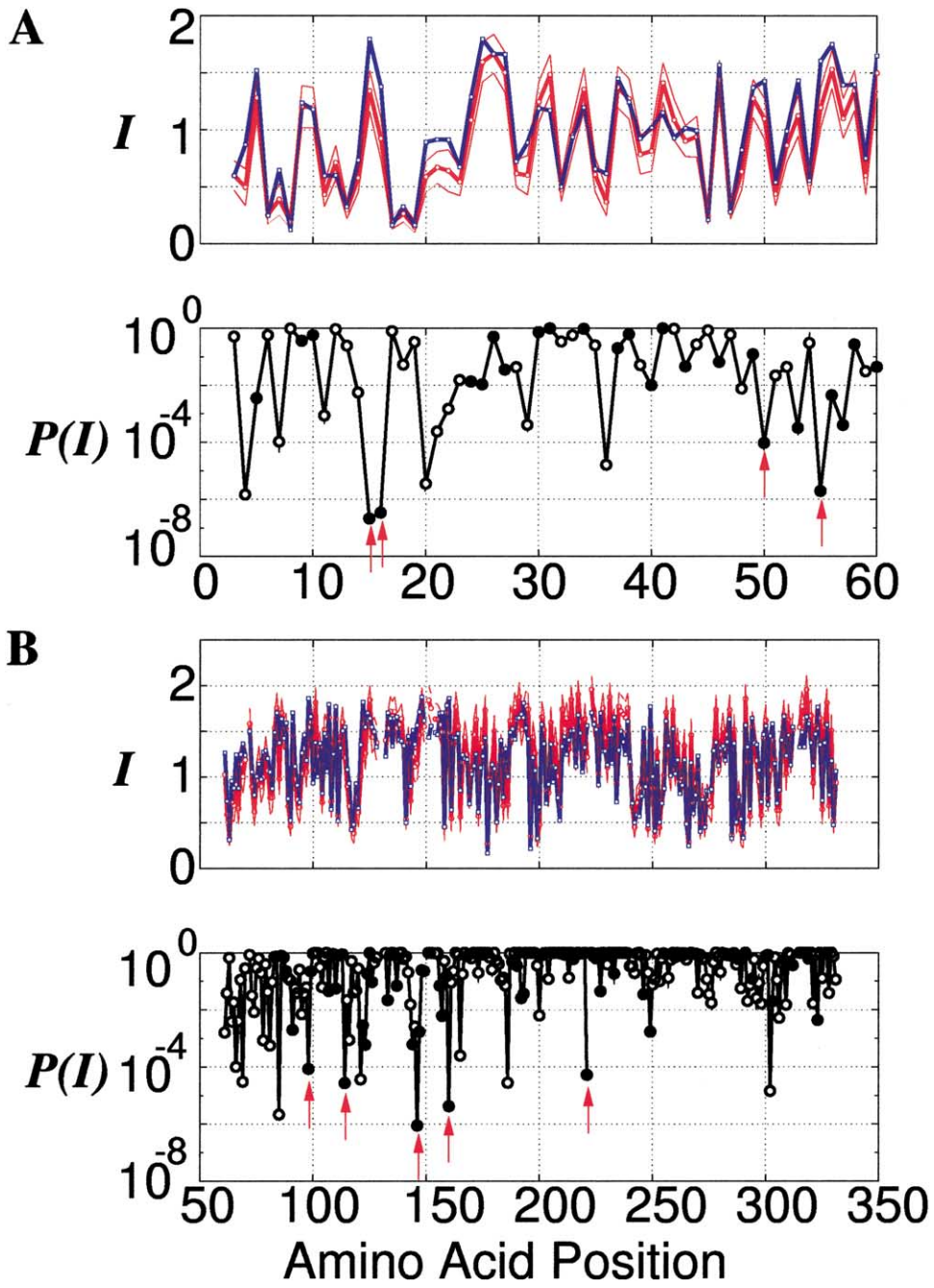

Figure 1. Observed I (blue) and the mean expected I^{exp} (thick red) mutual information in DNA-binding (a) and ligand-binding (b) domains of the LacI family. Thin red lines show $I^{exp} \pm 2\sigma(I^{exp})$. $P(I)$ is statistical significance of mutual information. Filled circles indicated residues with $I > 1.0$. Positions with filled circles and low $P(I)$ are predicted specificity determinants. The number along the sequence are according to 1wet PDB structure.

is available, since very few amino acid residues provide specific recognition (see below). Extensive site-directed mutagenesis is used to find such residues, though frequently complicated by a need to discriminate between specific and non-specific effects of a mutation. Computational prediction of the specificity determinants can substantially reduce experimental efforts and provide guidance for rational re-design of protein function.[16,17]

Our method relies on the above assumption that binding specificity is conserved among orthologous proteins and is different in paralogous proteins. The idea of our method is (1) to start from a family of paralogs in one genome, find orthologs for each member of the family in other genomes and (2) identify residues that can better discriminate between these orthologous (specificity) groups.

Assumption of specificity conserved among orthologs is not necessarily true.[18] However, mislabeling of orthologs and paralogs and errors in specificity assignments may mask some

Table 1. Lists of specificity-determining residues as predicted by different methods

Method	DNA-binding domain	Ligand-binding domain
Model 1, equation (3)	$P < 10^{-5}$ $I > 1.0$: 15, 16, 50, 55	$P < 10^{-5}$ $I > 1.5$: 98, 114, 122, 146, 147, 160, 221, 249
Model 1, equation (4)	$P < 10^{-4}$ $I > 1.0$: 15, 16, 50, 55	$P < 10^{-4}$ $I > 1.5$: 98, 114, 146, 160, 221
Model 2	$P < 10^{-2}$: 15, 16, 55	$P < 10^{-2}$: 85, 98, 122, 146, 160, 221, 246, 249

Numbering is according to 1wet PDB file.

specificity-determining positions, but will not lead to spurious predictions of specificity determinants. Thus any noise decreases sensitivity of the method, but does not lead to appearance of false positives. In the case considered here, the orthology relationships were simple to resolve. Given the increasing amount of genomics data and the emergence of genome analysis techniques such as positional clustering and regulon identification, the possibility to analyze more complicated cases will increase steadily.

In its second part the method is similar to techniques of hierarchical analysis of residue conservation,[19] PCA in the sequence space,[20] evolutionary trace analysis[21,22] and prediction of functional sub-types.[23] All these techniques use multiple sequence alignment (MSA) to group proteins into sub-groups based on sequence similarity and then identify residues that confer the unique features of each sub-group. Lapidot *et al*.[24] compared the variability of positions in aligned olfactory receptors of human and mouse, and identified positions conserved in orthologs, but varying in paralogs. A complementary structure-based approach was developed by Johnson & Church to predict protein function using a prior knowledge of the binding-site residues.[25] In contrast to other methods, our method relies on the definition of sub-families based on gene orthology and a rigorous statistical procedure to predict specificity-determining residues. Our statistical procedure determines whether positions in the MSA can discriminate between functional sub-families better than the sequence similarity. Residues that satisfy these criteria are predicted to be specificity-determining. Primarily, our method does not require the knowledge of the protein structure and can tolerate certain substitutions within a sub-family.

Here we present results of our analysis applied to the LacI/PurR family of bacterial transcription factors. The main result of this study is that among 12 identified specificity-determining residues, three are binding the DNA and eight are binding the ligand in the ligand-binding domain. The available experimental information supports the critical role of the identified DNA-binding residues in determining the specificity of the DNA recognition. Analysis developed here is not limited to DNA-binding proteins and can be applied to any family of proteins where the clear orthology or functional grouping can be established.

Results

Specificity determinants of the LacI family

We have chosen the LacI family for our analysis because (1) it is one of the largest families of bacterial transcription factors, (2) the availability of complete bacterial genomes has allowed us to resolve orthology by positional analysis (see Materials and Methods), and (3) available experimental[26–28] and structural[29,30] information can be used to assess our predictions.

Figure 1 presents the mutual information I_i, the expected mutual information $I^{\exp}$ and the probability $P(I)$ computed for the LacI family using Model 1. Model 2 produces very similar results. This plot reveals several important features: First, it shows high correlation $\rho = 0.97$ between I_i and $I_i^{\exp}$. Very good agreement between I_i and $I_i^{\exp}$ demonstrates that the statistical model used to compute $I^{\exp}$ succeeded in explaining $\rho^2 = 94\%$ of variation in mutual information and is able to reproduce naturally higher mutual information due to high intra-family similarity of orthologs (see Methods). Second, the vast majority of amino acid residues in the LacI family exhibit weak association with the specificity as indicated by $P(I) \approx 1$. Third, very few positions have both low $P(I_i)$ and high I_i (shown by arrows in Figure 1). Amino acid residues in these positions have strong association with functional grouping (stronger than sequences on average), indicating the role of these positions in determining different specificities of different groups of orthologs.

Table 1 presents predicted specificity-determining amino acid residues. Importantly, although methods to estimate statistical significance are very different, sets of residues found by them are very similar. The specificity determinants are: 15, 16, 50 and 55, in the first domain; and 98, 114, 122, 146, 147, 160, 221 and 249 in the second domain (here and below the numbering is according to PurR; the PDB code 1wet).

Table 2 shows the pattern of conservation of predicted specificity determinants. As expected, most of these residues are conserved within orthologous groups and are different between different groups. Importantly, there are some exceptions from this rule in all specificity-determining positions (see Discussion).

To better understand the role of specificity-determining residues we map them onto the structures

Table 2. Specificity determinants in the proteins of PurR/LacI family

Residue	15	160	146	55	98	16	114	221	249	122	50	147
$P(I)$	3×10^{-11}	2×10^{-10}	4×10^{-9}	6×10^{-9}	1×10^{-8}	2×10^{-8}	2×10^{-7}	6×10^{-7}	4×10^{-6}	4×10^{-6}	6×10^{-6}	9×10^{-6}
I	1.79	1.86	1.57	1.60	1.86	1.38	1.55	1.56	1.77	1.82	1.43	1.71
AraR	P	V	N	I	N	H	H	S	E	E	Q	A
	P	L	N	V	N	D	Q	T	E	E	Q	A
KdgR	K	T	D	N	N	T	L	T	W	Q	V	S
	K	T	D	N	N	T	L	T	W	Q	A	S
CcpA	M	I	G	A	D	A	K	Y	E	M	V	T
	M	I	G	A	D	A	K	Y	E	M	V	T
	M	I	A	A	D	A	K	Y	E	M	V	S
	I	I	E	K	D	A	K	L	E	M	V	T
	M	I	A	A	D	A	K	Y	E	M	V	S
	M	I	G	A	D	A	K	E	E	L	V	T
	M	I	G	A	D	A	K	Y	E	M	V	T
	M	I	G	A	D	A	K	Y	E	M	V	A
	M	I	G	A	D	A	K	Y	E	M	V	S
	M	I	G	A	D	A	K	P	E	M	V	S
	M	I	G	A	D	A	K	Y	T	M	V	T
	M	I	A	A	D	A	K	Y	E	L	V	S
DegA	.	S	D	Q	E	.	N	S	L	S	L	R
	P	S	D	Q	E	T	N	S	L	S	L	R
	P	S	D	Q	E	T	N	S	L	G	L	R
	P	S	D	Q	E	T	N	S	L	G	L	R
YjmH	H	A	N	V	I	T	K	Y	D	V	N	R
	H	S	N	A	I	T	M	Y	D	M	N	R
RbsR	T	D	D	K	E	S	K	F	M	M	L	W
	T	D	D	K	E	S	K	F	A	L	L	W
	T	E	D	K	G	S	K	F	T	M	V	W
	F	I	D	K	D	T	K	F	M	A	V	R
PurR	T	D	D	K	H	T	K	F	I	M	V	W
	T	D	D	K	W	T	K	F	I	M	V	W
	T	D	D	K	K	T	K	F	V	M	V	W
	T	D	D	K	G	T	K	F	T	M	V	W
CytR	T	I	A	K	A	A	K	F	V	L	L	N
	T	I	A	K	A	A	K	F	V	L	M	N
	T	I	A	R	G	A	K	F	T	L	L	C
GalS	V	L	I	A	Y	A	K	P	S	H	N	N
	V	L	I	A	Y	A	Q	P	N	H	N	N
	V	L	I	A	Y	A	H	P	S	H	N	N
	V	L	I	A	Y	A	H	P	S	H	N	N
AscG	R	F	L	A	H	S	L	Y	E	H	I	D
	R	F	L	A	R	A	L	Y	E	H	A	D
	K	L	N	A	K	A	Q	N	D	Y	L	R
	K	C	N	S	K	A	L	W	D	Y	L	R
LacI	Y	F	D	V	E	Q	Q	W	Q	N	L	V
	Y	F	D	A	R	Q	Q	W	Q	N	V	V
TreR	K	Y	A	R	Q	S	R	L	T	F	S	R
	K	Y	A	R	Q	S	R	L	T	F	S	R

(*continued*)

Table 2 Continued

Residue	15	160	146	55	98	16	114	221	249	122	50	147
$P(I)$	3×10^{-11}	2×10^{-10}	4×10^{-9}	6×10^{-9}	1×10^{-8}	2×10^{-8}	2×10^{-7}	6×10^{-7}	4×10^{-6}	4×10^{-6}	6×10^{-6}	9×10^{-6}
I	1.79	1.86	1.57	1.60	1.86	1.38	1.55	1.56	1.77	1.82	1.43	1.71
	K	Y	A	R	Q	S	R	L	T	F	S	M
GntR	K	F	M	S	G	M	Y	S	D	S	A	D
	K	F	M	S	G	M	W	S	D	T	A	D
	K	M	M	S	G	M	Y	S	D	T	A	E
IdnR	K	F	M	L	N	M	C	S	D	T	E	E
	K	F	M	L	N	M	Y	S	D	S	A	D
FruR	·	A	D	R	E	.	R	Y	Q	S	V	R
	R	A	D	R	E	T	R	Y	A	S	V	R
	K	E	D	R	D	T	R	F	T	A	A	R

I, mutual information; $P(I)$, statistical significance of mutual information. Proteins are grouped by paralogous specificity as indicated in the first column. Residues are sorted by $P(I)$.

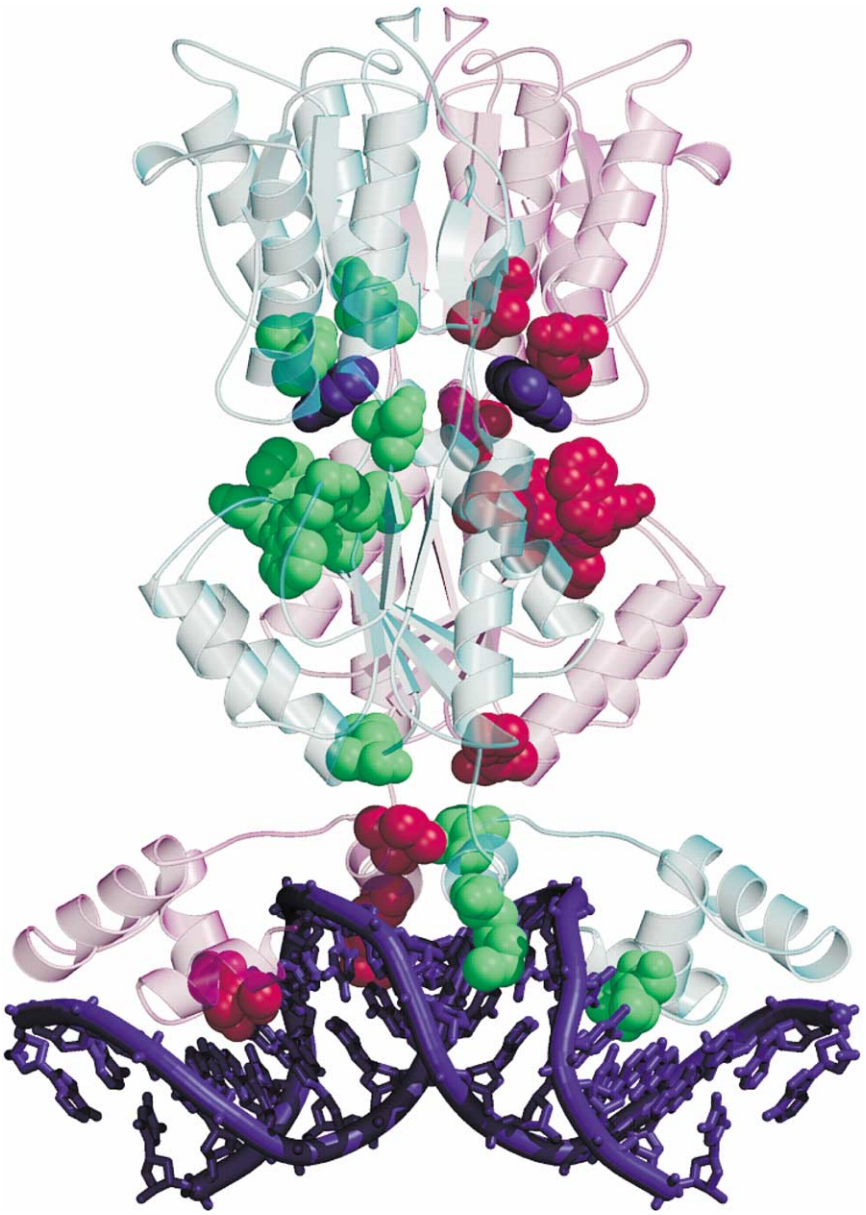

Figure 2. Structure of PurR bound to the DNA. Two chains of the dimer are shown semi-transparent in light green and pink. Predicted specificity determinants are shown by space-filling and colored red in the pink chain and green in the light green chain. The ligand (guanine) and the DNA are shown in blue. Notice deep penetration of some specificity-determining residues into the DNA and formation of the ligand-binding pocket by most of the others.

of the PurR and LacI–DNA complexes. Figure 2 shows the structure of the PurR–DNA complex with specificity-determining residues shown by space-filling atomic models with atoms of van der Waals radii. Clearly, these residues form two clusters in the structure: one around the DNA and other around the ligand. This result comes as no surprise, since proteins of the LacI family act as transcription repressors (activators) upon presence or absence of particular small molecules (sugars, nucleotides, etc.). Hence, paralogous proteins differ in specificity of both DNA and small molecule (ligand) recognition. The two identified spatial clusters supposedly determine this specificity.

Examination of the structure brings us to the following conclusions. (1) First four specificity-determining residues in PurR Thr15, Thr16, Val50 and Lys55 (Tyr17, Gln18, Val52 and ALA57 in LacI) are located in the DNA-binding domain. Three of them (15, 16 and 55 in PurR; 17, 18 and 57 in LacI) are deeply buried in the DNA grooves forming a dense network of interactions with the bases (see Figure 3(d) and (e)). Val50 (Val52 in LacI) forms a hydrophobic contact with its counterpart on the other chain. (2) Six more specificity-determining residues (out of eight) Met122, Asp146, Trp147, Asp160, Phe221 and Ile249 (Asn125, Asp149, Val150, Phe161, Trp220 and Gln248 in LacI) are located in the ligand-binding

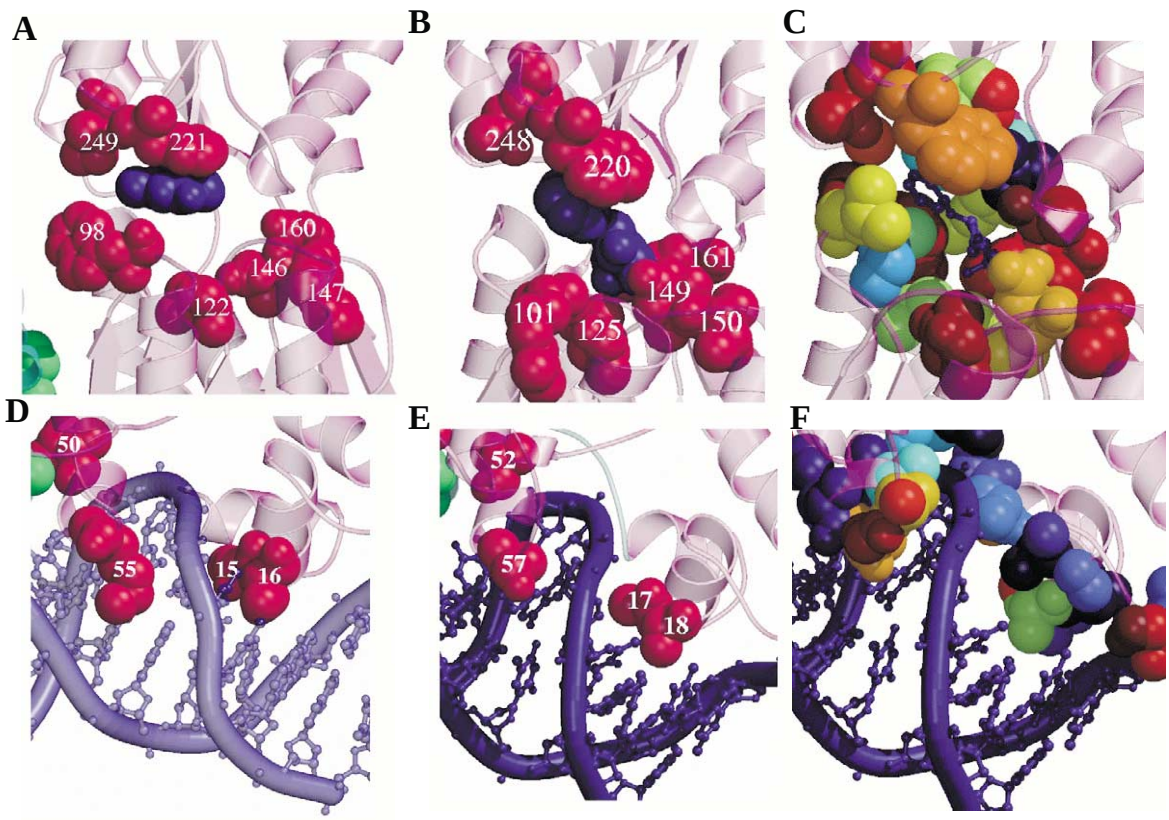

Figure 3. Detailed picture of the ligand binding pockets ((a)–(c)) and protein–DNA interface ((d)–(f)) in PurR ((a) and (d)) and LacI ((b), (c), (e) and (f)). Predicted specificity determinants are shown in space-fill on (a), (b), (d) and (e). (c) and (f) represent residues at 6 Å proximity from the ligand (c) and DNA (f). These residues are colored by their conservation sequence entropy (S) in LacI family from most conserved in blue to most variable in red. Note that it is impossible to predict specificity determinant by proximity or conservation.

pocket. Five of them (Met122, Asp146, Asp160, Phe221 and Ile249) are within 8 Å from the ligand in PurR and within 5 Å in LacI (Asn125, Asp149, Phe161, Trp220 and Gln248) (see Figure 3(a) and (b)). The observed clustering of the identified amino acid residues around the ligand is striking, since the structure of the protein was not used in our analysis.

Such structural location indicates that identified residues are indeed involved in the specific recognition. While the DNA-binding residues determine motifs recognized on the DNA, the residues located close to the ligand determine the ligand-binding specificity of the protein. Since different orthologs have different ligands, these residues change from sub-family to sub-family, but stay the same within most sub-families. Phe221 in PurR and corresponding Trp220 in LacI are of a special interest as their aromatic rings directly interact with aromatic ligands. Two other residues (Trp98 and Lys114 in PurR; Arg101 and Gln117 in LacI) do not belong to either of the clusters, as they are located far from the DNA and the ligand. They either are "false positives", or have some special role in the allosteric regulation.[31] Indeed, Val50, Trp98 and Lys114 of one chain interact tightly with the other chain, specifically, Val50 and Lys114 interact with side-chains of Lys114 and Val50 from the other chain. These residues can be important for correct dimerization and hence exhibit sought covariation with functional grouping. In summary, the structural location of identified residues supports our findings that they serve as specificity determinants in proteins of the LacI family. This includes the specificity of the DNA recognition and the ligand-binding specificity.

Although putative specificity determinants are located close to the DNA and the ligand, few residues that are contacting the ligand or the DNA exhibit the conservation pattern of specificity determinants. Figure 3(c) and (f) shows residues located within 6 Å from the DNA (Figure 3(f)) and the ligand (Figure 3(c)). The residues are colored by their sequence entropy, S, which is traditionally used to estimate the degree of evolutionary conservation.[32]

This picture demonstrates that specificity determinants constitute only a small fraction of all residues located on binding interfaces. It is also clear that the value of sequence entropy cannot discriminate specificity determinants from other residues. There observations emphasize challenges in identifying specificity determinants by traditional evolutionary and structural analysis.

Discussion

In this study we suggested a method to identify specificity-determining residues in proteins. We applied it to one of the largest family of bacterial transcription factors and obtained a set of putative specificity-determining residues. Mapping of these residues onto a protein structure showed that most of identified residues belong to two spatial clusters. Residues of one cluster bind the DNA, while residues of the other cluster form a ligand pocket of the protein. This finding is consistent with the function of transcription factors of this family: they repress transcription by binding the DNA and release transcription when a particular ligand is present. (Conversely, some proteins in the family, e.g. PurR, bind the DNA only when the ligand is present.) Paralogous proteins of this family differ from each other in the ligands they recognize and in the DNA sites they bind. Hence, two clusters of residues found by our method presumably determine specificity of these two recognition processes.

Our analysis suggested residues 15, 16 and 55 as primary determinants of the DNA-binding specificity. The role of positions 15, 16 and 55 in specific DNA recognition is evident from a series of mutant experiments.[26,27,28] Extensive site-directed mutagenesis of the second helix of LacI showed that residues 15 and 16 are essential for DNA-binding specificity.[33] When Tyr15 and Gln16 of LacI were mutated to the residues present in these positions in the paralogs (MalI, RafR, CytR, etc.) the mutants were preferentially binding operators of the respective paralogs. Similarly, when GalR was mutated to have the residues of LacI in positions 15 and 16, mutant GalR was specifically binding sites of LacI.[26,27] These experiments strongly support our result that positions 15 and 16 are responsible for determining DNA-binding specificity in proteins of the LacI family. Another residue found by our analysis is residue 55. Although residue 55 is binding DNA in the minor groove, this residue was shown to be critical for the DNA recognition by PurR.[28] Our results suggest that a triple mutant (15, 16 and 55) should have a higher specificity and affinity to paralogous operators.

To the best of our knowledge, residues identified in the ligand-binding domain (except for 146) have not yet been the subjects of protein engineering studies. Although mutations of several other residues were shown to interfere with the ligand binding, it is not clear how they influence specificity (as opposed to affinity) of the ligand recognition. Most of mutations in the region were shown to drastically reduce the affinity. Our analysis suggests ways to do rational re-design of the ligand binding specificity. One can "transplant" some or all of the outlined residues from a paralog to LacI and measure the mutant's binding constants for various ligands normally bound by this paralog. The main question posed by our study is whether the specificity can be re-designed by changing a small set of the predicted residues.

Another possible application of the putative specificity determinants is in more focused prediction of the DNA-binding specificity. Instead of considering all interactions between the DNA and the protein, one can focus on the interactions formed by the specificity determinants. This approach is a subject of our current research.

The most important part of the presented algorithm is the procedure used to calculate statistical significance of the mutual information. Specificity determinants were selected as residues having both high I and very low $P(I)$. Note, that selection by high I alone would yield a very different (and very large) set of residues (see Figure 1, filled circles). Most of such residues do not have statistically significant association with grouping ($P(I) \approx 1$). This observation emphasizes the importance of statistical test in our analysis.

Although promising, our analysis has its limitations. It relies heavily on the grouping of proteins by orthology. To resolve orthology, one needs to have (almost) complete genomes of several closely related organisms. This makes our analysis significantly data demanding. Even if complete genomes are available, orthology may not be easily resolved when very similar paralogs are present or when genomes are too diverged from each other. Our analysis also assumes that orthologs have the same function and specificity. This is likely true for evolutionary close organisms, where orthologs had not enough evolutionary time to diverge in specificity. One way to avoid these pitfalls is to use proteins where conserved specificity has been experimentally verified or confirmed by independent genomic positional or regulatory analysis. Using genomicly resolved orthologs one has to rely on the statistical significance, $P(I)$. If a dataset is "contaminated" with false orthologs or orthologs with diverged specificity, no residues would have low $P(I)$.

Our analysis also relies on the assumption that the same residues determine specificity of paralogs. Little is known about the spatial location of the specificity determinants. Active site residues, however, are known to have very conserved spatial location in the families of homologous proteins. Active sites have the same spatial location, even when similarity between the sequences is as low as 10%. For example, proteins of the TIM-barrel and flavodoxin folds have "super-sites", i.e. active sites in the same spatial location, although amino acid residues forming the sites and the biochemical function of these proteins have widely diverged.[34,35] This extreme conservation of the spatial location of functionally crucial residues supports the assumption of common location of the specificity determinants. Since most orthologs and close paralogs have high sequence similarity, residues matched in the MSAs are likely to have common spatial location. However, recognition of molecules of various shapes (ligands, protein

interfaces) may involve different interactions and therefore variable spatial location of the specificity residues cannot be ruled out. The most direct test of our assumption would be to perform the experiments suggested above. Calculation of statistical significance also constitutes an internal test of our method. If sets of residues which determine specificities of paralogs differ, our method will identify an overlap between these sets as having low $P(I)$. If the sets do not overlap, no residues would have substantially low $P(I)$.

As can be seen from Table 2 our method, in contrast to evolutionary trace analysis,[21] can tolerate certain substitutions within a group of orthologs. All amino acid residues, however, are assumed to be equally distinct in their properties. In other words, substitutions $I \rightarrow L$ and $I \rightarrow H$ are treated equally, while in reality the change of the physical properties of amino acid residues depends on the type of the substitutions. We are currently developing a method to identify specificity determinants with different spatial locations, which will also take into account physical properties of amino acid residues.

Here we have suggested a method to find residues that determine the specificity of the protein recognition. The method is based on discrimination between orthologous and paralogous proteins, taking advantage of several complete bacterial genomes to identify them. The method does not require a solved 3D structure of a protein to predict specificity determinants. Analysis of a large LacI family of bacterial transcription factors found two groups of residues as the putative DNA-binding and ligand-binding specificity determinants. Predictions of the DNA-binding residues are strongly supported by the earlier experimental results. Results of our analysis suggest targeted protein engineering experiments aimed at rational re-design of the protein specificity.

Materials and Methods

The key idea of this method is to compare paralogous and orthologous proteins from the same family. As a rule, all paralogous and orthologous proteins have the same biochemical function. Paralogous proteins, however, usually have different specificity as they act on different targets, e.g. bind different ligand or different sites on the DNA. Orthologous proteins, in contrast, have the same specificity in different organisms, e.g. bind the same ligand and similar DNA sites in related genomes. Hence, orthologous proteins carry the same or similar specificity-determining residues, whereas paralogous proteins carry different ones. On the basis of this idea, our analysis is looking for residues that are conserved among orthologs and different in paralogs. More generally, we are looking for residues that can discriminate between different paralogs, while grouping orthologs together. We call these residues specificity determining.

The analysis works as following: First, in a group of homologous proteins, paralogs from one organism are selected. Second, for each of the paralogs we find its orthologs in related organisms and build a MSA using ClustalW.[36] Third, we compute the mutual information for each position of the MSA. The mutual information determines how well a residue in the MSA can discriminate between orthologous groups. The forth, and the most important step is to compute the statistical significance of the discrimination and to select residues that can discriminate significantly better than the others. These residues are the specificity determinants.

Selection of orthologs

A list of complete and almost complete bacterial genomes used in this study and a full list of orthologs is provided below. Homologs of LacI and PurR of *Escherichia coli* were identified using GenomeExplorer[37] and supplemented by proteins from SwissProt.[38] Then phylogenetic trees were constructed using the neighbor-joining procedure implemented in ClustalW.[36]

Only unambiguous groups of orthologs were selected and identified by (1) absence of duplications in the corresponding sub-branches of the tree (in two cases duplications in one genome were allowed where the proteins were known to have the same ligand and DNA-binding specificity); (2) coinciding functional annotation when known (no proteins with different specificity were included in one group) and (3) genomic positional analysis (genes encoding candidate orthologs should belong to orthologous loci, that is, have orthologous neighbors).

Mutual information

To identify residues that can discriminate between paralogous proteins (different specificity), merging orthologs (same specificity) together we use the mutual information as a measure of association with the specificity. Mutual information is frequently used in computational biology for co-variational analysis in RNA and proteins.[39,40]

If $x = 1, \ldots, 20$ is a residue type, $y = 1, \ldots, Y$ is the specificity index which is the same for all proteins of the same specificity group and is different for different groups, and Y is the total number of specificity groups, then the mutual information at position i of the MSA is:

$$I_i = \sum_{\substack{x=1,\ldots,20 \\ y=1,\ldots,Y}} f_i(x,y) \log \frac{f_i(x,y)}{f_i(x) f(y)} \qquad (1)$$

where $f_i(x)$ is the frequency of residue type x in position i of the MSA, $f(y)$ is the fraction of proteins belonging to the group y, and $f_i(x,y)$ is the frequency of residue type x in the group y at position i. Mutual information has several important properties: (1) it is non-negative; (2) it equals zero if and only if x and y are statistically independent; and (3) a large value of I_i indicates a strong association between x and y.[41] Unfortunately, a small sample size and a biased composition of each column in the MSA influences I_i a lot. For example, positions with less conserved residues tend to have higher mutual information. Hence, we cannot rely on the value of I_i as an indicator of specificity association, instead we estimate the statistical significance of I_i.

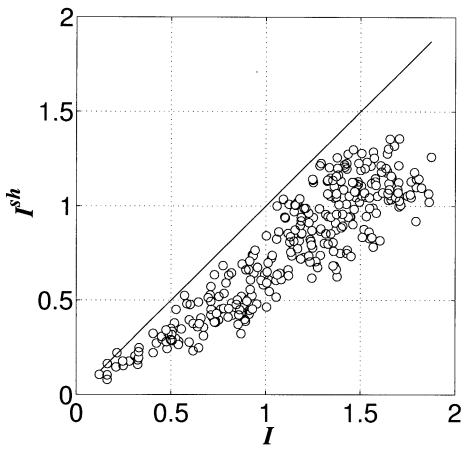

Figure 4. Observed mutual information I and mutual information I^{sh} obtained for amino acid residues shuffled in each column of the MSA (circles). The correlation coefficient between I and I^{sh} is 0.97. Note that I^{sh} is systematically lower than I due to closer relations between orthologs than paralogs. The line shows $x = y$.

Statistical significance

Since mutual information can be biased due to the small sample size or biased amino acid composition, we cannot rely on the value of mutual information to identify the specificity determinants. Instead, we compute the statistical significance $P(I)$ of the mutual information and use it together with I to predict the specificity-determining residues. Calculation of statistical significance is the most important component of the method. We present two different approaches, which, however, produce very similar results.

A standard way of computing $P(I)$ is by shuffling.[42] However, this method is unacceptable for the following reason. Naturally, proteins within each specificity group (orthologs) are much more similar to each other than proteins from different groups (paralogs). Hence, amino acid residues at every position are somewhat associated with the functional grouping, producing I higher than the mutual information obtained by shuffling (see Figure 4). Developing a statistical test we have to take into account the naturally higher similarity between orthologs in comparison to paralogs. In other words, we need a statistical test to identify positions that are stronger associated with the functional grouping, than the whole proteins on average.

To accomplish this task, we developed two statistical models. The first model uses linear transformation to take into account a bias introduced by higher intra-group similarity of orthologs. The second model simulates the actual evolutionary process of duplication followed by further accumulation of mutations.

Model 1

In outline, we first compute I^{sh} using shuffling, then transform it to I^{exp} using the maximum likelihood estimator to take into account higher similarity between orthologs and finally compute the desired statistical significance $P(I)$.

We need to take into account the fact that orthologs are more closely related than paralogs. Due to this fact, sequence similarity between orthologs is higher than between paralogs (see Figure 5(a)). As a result, any position in the MSA has certain association with grouping by orthology. Specificity determinants, however, must have stronger association with this grouping than any position on average. To compute $P(I)$ we start from a null-hypothesis that amino acid residues in all positions of the MSA have the same association with grouping by orthology.

Consider the MSA a_i^m, $i = 1, \ldots, L$, $m = 1, \ldots, M$, where a_i^m is the residue in position i of the mth protein, L is the length of the alignment and M is the total number of aligned proteins. For each position i we take a column $\mathbf{a}_i$ of the MSA and randomly shuffle this column. Next we compute the mutual information of the shuffled ($\mathbf{a}_i^{sh}$) grouping: $I^{sh} = I(\mathbf{a}^{sh})$. This procedure is repeated 10^4 times to get the distribution of the mutual information for a shuffled column.

As explained above, I^{sh} is systematically lower than I due to the higher intra-group similarity between orthologs. We assume that the systematic bias introduced by this similarity is position-independent and linear. Figure 4 shows I^{sh} *versus* I for each position. Linear correlation coefficient of 0.97 and the lack of noticeable non-linear trends justify the use of linear transformation for I^{sh}.

To compute expected mutual information I^{exp} we make transformation:

$$I_i^{exp} = \alpha I_i^{sh} + \beta \tag{2}$$

Note that parameters α and β do not depend on i. α and β are obtained by either minimizing the deviation between the observed and mean of expected mutual information:

$$\mathscr{D} = \sum_i (I_i - \langle I_i^{exp} \rangle)^2 = \sum_i (I_i - \alpha \langle I_i^{sh} \rangle - \beta)^2 \tag{3}$$

or by a maximal likelihood estimator which maximizes the likelihood of observed mutual information:

$$\mathscr{L} = \prod_i P(I_i) \sim \sum_i \left(\frac{I_i - \langle I_i^{exp} \rangle}{\sigma(I_i^{exp})} \right)^2 \sim \sum_i \left(\frac{I_i - \alpha \langle I_i^{sh} \rangle - \beta}{\alpha \sigma(I_i^{sh})} \right)^2 \tag{4}$$

$\langle I_i^{sh} \rangle$ and $\sigma(I_i^{sh})$ are the mean and the variance of I_i^{sh} obtained by 10^4 random shufflings. The later equation is obtained assuming normal distribution of $I - I^{exp}$, and analysis of the residuals shows that it is in fact normal even at very large deviations from the mean (checked by 10^5 shufflings, data not shown). To obtain α and β using equation (3) we make linear regression $I_i = \alpha \langle I_i^{sh} \rangle + \beta$. To obtain α and β using equation (4) we use reverse regression:

$$\frac{\langle I_i^{sh} \rangle}{\sigma(I_i^{sh})} = A \frac{I_i}{\sigma(I_i^{sh})} + B \frac{1}{\sigma(I_i^{sh})})$$

and then obtain $\alpha = 1/A$ and $\beta = -B/A$. To avoid possible bias introduced by irrelevant positions with low I_i, we compute α and β using positions with $I_i > 0.5$.

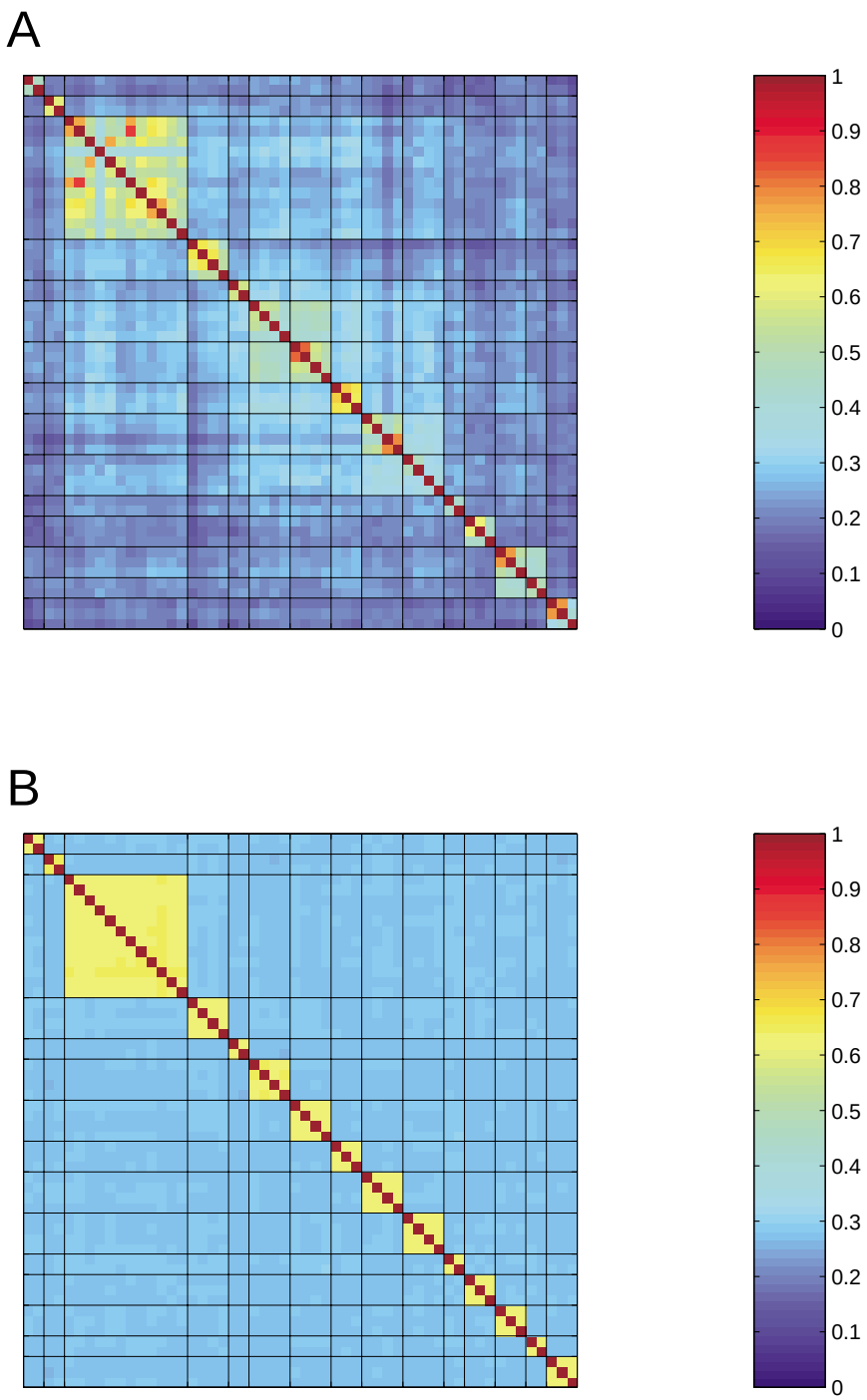

Figure 5. Sequence identity between the proteins of LacI family. Color diagram of the fraction of identical residues (sequence identity) between all pairs of sequences, i.e. cell (i, j) shows the identify between sequences i and j of the MSA. (a) Continuous lines separate orthologous groups. Notice higher sequence identity within the groups (near-diagonal squares), than between them (blue background). (b) Sequence identity between pseudo-random sequences generated to compute $P(I)$ capturing higher intra-group identity. Averaged over 20 sets of pseudo-random sequences.

After α and β are computed, we obtain the desired probability:

$$P_i = P(I_i) = \frac{1}{\sqrt{2\pi\sigma}(I_i^{\exp})} \int_{I_i}^{+\infty} \exp\left(-\frac{(I - \langle I_i^{\exp}\rangle)^2}{2\sigma^2(I_i^{\exp})}\right) dI$$

$$= \frac{1}{\sqrt{2\pi}} \int_{Z_i}^{+\infty} \exp\left(-\frac{z^2}{2}\right) dz$$

where:

$$Z_i = \frac{I_i - \langle I_i^{\exp}\rangle}{\sigma(I_i^{\exp})} \quad (5)$$

Very low P_i indicates that the null-hypothesis does not hold for position i and residues in this position are in fact stronger associated with the specificity grouping than the whole proteins. Thus positions in the MSA that exhibit low P_i and high I_i are the specificity determinants.

$I^{\exp}$ and P_i obtained using either equations (3) or (4) lead to very similar results (see Table 1). However, the MLE (4) is a more general way of deriving parameters of the model and in our analysis we relied on $I^{\exp}$ and P_i obtained this way.

Model 2

This model does not utilize shuffling to compute $I^{\exp}$. Instead we model evolution of the protein family and

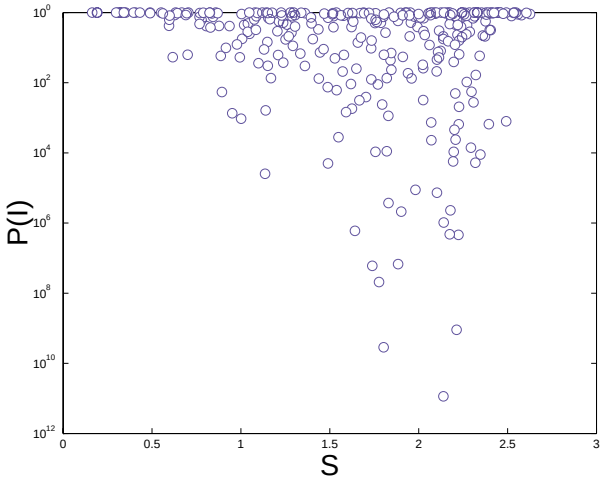

Figure 6. Statistical significance $P(I)$ *versus* sequence entropy S for each position in the MSA. Sequence entropy indicates the rate of amino acid residue substitution. Specificity determinants (low $P(I)$) have high substitution rate, but not *vice versa*.

generate a set of pseudo-random protein sequences such that (i) sequence entropy S_i (see below) of each column in the MSA of pseudo-random proteins is preserved; (ii) generated sequences have intra- and inter-group similarity similar to those of the studied proteins. These simulations explicitly take into account the fact that orthologs are more closely related than paralogs. Using obtained pseudo-random proteins we compute mutual information I^{rnd}. Finally, we compute $P_i = P(I_i)$ as the probability of observing mutual information above I_i for the pseudo-random proteins.

We start from a null-hypothesis that all positions in the MSA have the same association with specificity grouping. To compute P_i, we need to generate sequences that have the same intra-group and inter-group similarity as the orthologs and the paralogs, respectively. This is achieved by simulating evolution of these proteins in the following manner:

(i) Generate a "parent" sequence $\mathbf{b}^y = b_i^y$, $i = 1, \ldots, L$ for each group of orthologs $y = 1, \ldots, Y$. An amino acid residue b_i^y is generated randomly from the distribution of amino acids at position i $f_i(x)$. This step simulates evolution of paralogous proteins by duplication.

(ii) Generate a sequence of the mth protein c_i^m from the "parent" sequence of its group y. We assume that during speciation, that followed duplication, some amino acids did not get substituted. We simulate this by introducing the probability μ of inheriting an amino acid from the "parent" protein without a substitution. Hence $c_i^m = b_i^y$ with probability μ, and c_i^m is taken randomly from $f_i(x)$ with probability $1 - \mu$. This step simulates evolution of orthologs through speciation.

Parameter μ controls the intra-group similarity: for $\mu = 1$ all generated orthologs are identical, whereas for $\mu = 0$ they are as different as paralogs. We set $\mu = 0.85$ to get the intra-group similarity close to that of the studied natural proteins (see Figure 5).

After pseudo-random correlated sequences are generated, we compute I_i^{rnd}. for them using equation (1) The sequences (including "the parents") are generated in 10^3 independent runs yielding the distribution of the mutual information $f_i(I^{\text{rnd}})$. Assuming normal distribution of I^{rnd} we get P_i as:

$$P_i = P(I_i) = \frac{1}{\sqrt{2\pi}\sigma(I_i^{rnd})} \int_{I_i}^{+\infty} \exp\left(-\frac{(I - \langle I_i^{rnd}\rangle)^2}{2\sigma^2(I_i^{rnd})}\right) \mathrm{d}I$$

$$= \frac{1}{\sqrt{2\pi}} \int_{Z_i}^{+\infty} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z$$

where:

$$Z_i = \frac{I_i - \langle I_i^{\text{rnd}}\rangle}{\sigma(I_i^{\text{rnd}})} \tag{6}$$

Alternatively one can get P_i as the fraction of runs in which I_i^{rnd} exceeded I_i. Both estimates give very similar results, but equation (6) allows us to estimate P_i even when $I_i^{\text{rnd}} < I_i$ in all runs.

To make sure that Model 2 correctly reconstructs higher sequence similarity between orthologs, we calculated sequence identity between every pair of sequences in the LacI family and in the simulated pseudo-random proteins. Figure 5 shows these results. Clearly pseudo-random sequences have desired higher intra-group similarity.

Sequence entropy

The degree of sequence variability in position i of the MSA is measured by the sequence entropy:

$$S_i = -\sum_{x=1}^{20} f_i(x) \log f_i(x) \tag{7}$$

Sequence entropy can be used as an estimator of the evolutionary substitution rate. Figure 6 presents $P(I)$ as a function of S. The plot shows that most of the selected amino acid residues (low $P(I)$ points) have medium to high substitution rates. However, few positions with the medium substitution rate have low $P(I)$ (see Figure 3(c) and (f)).

List of organisms and proteins

List of complete and almost complete genomes used in this study: *Bacillus subtilis*[43] *Clostridium acetobutylicum*[44] *Streptococcus pyogenes*[45] *Streptococcus pneumoniae*[46] *Enterococcus faecalis* (DOE-JGI, www.jgi.doe.gov) *Vibrio cholerae*[47] *Escherichia coli*[48] *Yersinia pestis*[49] *Haemophilus influenzae*[50] *Pseudomonas aeruginosa*[51]

The list of orthologs includes: AraR from *B. subtilis* and *C. acetobutylicum*; KdgR from *S. pneumoniae* and *S. pyogenes*; CcpA from *B. subtilis*, *Bacillus megaterium*, *Staphylococcus aureus*, *Listeria monocytogenes*, *E. faecalis*, *S. pneumoniae*, *S. pyogenes*, *S. pneumoniae*, *Lactococcus lactis*, *Q48518_LACCA*, *Q9ZHP7_thE97*; DegA from *B. subtilis*, *E. faecalis*, *S. pneumoniae*, and *S. pyogenes*; YjmH from *B. subtilis* and *C. acetobutylicum*; RbsR from *E. coli*, *V. cholerae*, *H. influenzae*, and *P. aeruginosa*; PurR from *E. coli*, *Y. pestis*, *V. cholerae*, and *H. influenzae*; CytR from *E. coli*, *Y. pestis*, and *V. cholerae*; GalS/GalR *E. coli*, *Y. pestis*, and *H. influenzae*; AscG/AscG1 from *E. coli*, *Y. pestis*, and *V. cholerae*; LacI from *E. coli* and *Y. pestis*; TreR from *E. coli*, *Y. pestis*, and *V. cholerae*; GntR from *E. coli*, *Y. pestis*, and *V. cholerae*; FruR from *E. coli*, *Y. pestis*, and *V. cholerae*; and IdnR from *E. coli* and *Y. pestis*.

Acknowledgments

We are grateful to Alexander van Oudenaarden for helpful discussions and initiation of experimental work to test our predictions. We also acknowledge useful comments made by Richard Goldstein and Eugene Shakhnovich while discussing this work. L.M. is partially supported by William F. Milton Fund and John F. and Virginia B. Taplin Award. M.G. is partially supported by grants from INTAS (99-1476), Howard Hughes Medical Institute (55000309), and the Ludwig Cancer Research Institute. We are grateful to Dmitry Rodionov for the help with the data and Olga Laikova for useful discussions.

References

1. Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.
2. Fitch, W. (2000). Homology a personal view on some of the problems. *Trends Genet.* **16**, 227–231.
3. Koonin, E. (2001). An apology for orthologs—or brave new memes. *Genome Biol.* **2**, 1005.
4. Petsko, G. (2001). Homologuephobia. *Genome Biol.* **2**, 1002.
5. Jensen, R. (2001). Orthologs and paralogs—we need to get it right. *Genome Biol.* **2**, 1002.
6. Gerlt, J. & Babbitt, P. (2000). Can sequence determine function? *Genome Biol.* **1**, 5.
7. Makarova, K., Aravind, L., Galperin, M., Grishin, N., Tatusov, R., Wolf, Y. & Koonin, E. (1999). Comparative genomics of the archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res.* **9**, 608–628.
8. Tatusov, R., Galperin, M., Natale, D. & Koonin, E. (2000). The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucl. Acids Res.* **28**, 33–36.
9. Gelfand, M., Koonin, E. & Mironov, A. (2000). Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucl. Acids Res.* **28**, 695–705.
10. Gerlt, J. & Babbitt, P. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246.
11. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
12. Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F. *et al.* (2001). Regulondb (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* k-12. *Nucl. Acids Res.* **29**, 72–74.
13. McCue, L., Thompson, W., Carmack, C., Ryan, M., Liu, J., Derbyshire, V. & Lawrence, C. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucl. Acids Res.* **29**, 774–782.
14. Hertz, G. & Stormo, G. (1999). Identifying DNA and protein patterns with statistically significant a ligments of multiple sequences. *Bioinformatics*, **15**, 563–577.
15. Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J. & Stormo, G. (2001). A comparative genomics approach to prediction of new members of regulons. *Genome Res.* **11**, 566–584.
16. Fersht, A. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, WH Freeman & Co, San Francisco.
17. Ballinger, M., Tom, J. & Wells, J. (1996). Furilisin: a variant of subtilisin bpn′ engineered for cleaving tribasic substrates. *Biochemistry*, **35**, 13579–13585.
18. Theissen, G. (2002). Secret life of genes. *Nature*, **415**, 741.
19. Livingstone, C. & Barton, G. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756.
20. Casari, G., Sander, C. & Valencia, A. (1995). A method to predict functional residues in proteins. *Nature Struct. Biol.* **2**, 171–178.
21. Lichtarge, O., Bourne, H. & Cohen, F. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
22. Lichtarge, O., Yamamoto, K. & Cohen, F. (1997). Identification of functional surfaces of the zinc binding domains of intracellular receptors. *J. Mol. Biol.* **274**, 325–337.
23. Hannenhalli, S. & Russell, R. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303**, 61–76.
24. Lapidot, M., Pilpel, Y., Gilad, Y., Falcovitz, A., Sharon, D., Haaf, T. & Lancet, D. (2001). Mouse–human orthology relationships in an olfactory receptor gene cluster. *Genomics*, **71**, 296–306.
25. Johnson, J. & Church, G. (2000). Predicting ligand-binding function in families of bacterial receptors. *Proc. Natl Acad. Sci. USA*, **97**, 3965–3970.
26. Sartorius, J., Lehming, N., Kisters-Woike, B., von, W.-B. & Muller-Hill, B. (1991). The roles of residues 5 and 9 of the recognition helix of *lac* repressor in *lac* operator binding. *J. Mol. Biol.* **218**, 313–321.
27. Lehming, N., Sartorius, J., Kisters-Woike, B., von, W.-B. & Muller-Hill, B. (1990). Mutant *lac* repressors with new specificities hint at rules for protein–DNA recognition. *EMBO J.* **9**, 615–621. Published erratum appears in *EMBO J.* **9**, 1674.
28. Glasfeld, A., Koehler, A., Schumacher, M. & Brennan, R. (1999). The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. *J. Mol. Biol.* **291**, 347–361.
29. Schumacher, M., Glasfeld, A., Zalkin, H. & Brennan, R. (1997). The X-ray structure of the purr–guanine–purf operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated. *J. Biol. Chem.* **272**, 22648–22653.
30. Bell, C. & Lewis, M. (2001). Crystallographic analysis of lac repressor bound to natural operator o1. *J. Mol. Biol.* **312**, 921–926.
31. Lockless, S. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
32. Pei, J. & Grishin, N. (2001). Al2co: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
33. Sartorius, J., Lehming, N., Kisters, B., von, W.-B. & Muller-Hill, B. (1989). lac repressor mutants with double or triple exchanges in the recognition helix bind specifically to lac operator variants with multiple exchanges. *EMBO J.* **8**, 1265–1270.
34. Hasson, M., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G. *et al.* (1998). Evolution of an enzyme active site: the structure of a new crystal

form of muconate lactonizing enzyme compared with mandelate racemase and enolase. *Proc. Natl Acad. Sci. USA*, **95**, 10396–10401.
35. Russell, R., Sasieni, P. & Sternberg, M. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903–918.
36. Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
37. Mironov, A., Vinokurova, N. & Gelfand, M. (2000). Software for analysis of bacterial genomes. *Mol. Biol. (Mosk)*, **34**, 253–262.
38. Bairoch, A. & Apweiler, R. (2000). The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucl. Acids Res.* **28**, 45–48.
39. Clarke, N. (1995). Covariation of residues in the homeodomain sequence family. *Protein Sci.* **4**, 2269–2278.
40. Gorodkin, J., Staerfeldt, H., Lund, O. & Brunak, S. (1999). Matrixplot: visualizing sequence constraints. *Bioinformatics*, **15**, 769–770.
41. Cover, T. & Thomas, J. (1991). *Elements of Information Theory*, Wiley, New York.
42. Good, P. (1994). *Permutation Tests : A Practical Guide to Resampling Methods for Testing Hypotheses* Springer Series in Statistics, Springer, New York.
43. Kunst, F., Ogasawara, N., Moszer, I. *et al.* (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
44. Nolling, J., Breton, G., Omelchenko, M. *et al.* (2001). Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol.* **183**, 4823–4838.
45. Ferretti, J., McShan, W., Ajdic, D. *et al.* (2001). Complete genome sequence of an m1 strain of *Streptococcus pyogenes*. *Proc. Natl Acad. Sci. USA*, **98**, 4658–4663.
46. Tettelin, H., Nelson, K., Paulsen, I. *et al.* (2001). Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science*, **293**, 498–506.
47. Heidelberg, J., Eisen, J., Nelson, W. *et al.* (2000). DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.
48. Blattner, F., Plunkett, G., Bloch, C. *et al.* (1997). The complete genome sequence of *Escherichia coli* k-12. *Science*, **277**, 1453–1474.
49. Parkhill, J., Wren, B., Thomson, N. *et al.* (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–527.
50. Fleischmann, R., Adams, M., White, O. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science*, **269**, 496–512.
51. Stover, C., Pham, X., Erwin, A. *et al.* (2000). Complete genome sequence of *Pseudomonas aeruginosa* pa01, an opportunistic pathogen. *Nature*, **406**, 959–964.

Edited by J. Karn

(Received 13 February 2002; received in revised form 7 June 2002; accepted 12 June 2002)